DOCUMENT RESUME

ED 411 246 TM 027 112

AUTHOR Monaco, Malina
TITLE A Monte Carlo Study of Skewed Theta Distributions on DIF Indices.
PUB DATE 1997-03-00
NOTE 26p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, March 25-27, 1997).
PUB TYPE Numerical/Quantitative Data (110) -- Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Ability; Estimation (Mathematics); *Item Bias; *Monte Carlo Methods; *Statistical Distributions; Tables (Data); *Test Items
IDENTIFIERS Item Bias Detection; Mantel Haenszel Procedure; *Skew Curves; *Theta Estimates

ABSTRACT

The effects of skewed theta distributions on indices of differential item functioning (DIF) were studied, comparing Mantel Haenszel (N. Mantel and W. Haenszel, 1959) and DFIT (N. S. Raju, W. J. van der Linden, and P. F. Fleer) (noncompensatory DIF). The significance of the study is that in educational and psychological data, the distributions one deals with are sometimes not normal, and that could translate into a skewed distribution. Monte Carlo procedures were used to generate the data under normal and skewed conditions for 1,000 and 500 examinees each in focal and reference groups and a test length of 40 items. The thetas for the examinees were drawn from normal, positively skewed, and negatively skewed distributions. Item and ability parameters were estimated though the BILOG computer programs. The effectiveness of each DIF index was determined by examining the true positive ("hits") and the false positive rates. The Mantel Haenszel, Lord's chi square, and noncompensatory DIf detected uniform DIF items very well. Precision of detection increased as sample size increased with all methods. All three methods tended to have difficulty detecting nonuniform DIF items. Precision declined for all methods at the extreme skew, but moderate skews do not affect detection much at all. It is recommended that researchers use a combination of DIF detection methods as seems appropriate to the conditions. (Contains 11 tables and 23 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

A Monte Carlo Study of Skewed Theta Distributions on DIF Indices.

ED 411 246

Malina Monaco

Georgia State University

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Malina Monaco

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper presented at the annual meeting of the National Council on Measurement in Education
Chicago, March, 1997

Address Correspondence To:

Malina Monaco
Dept. of Educational Policy Studies
University Plaza
Georgia State University
Atlanta, GA 30303-3083

or

Send email To:

mmonaco@gsu.edu

TM027112

BEST COPY AVAILABLE

A Monte Carlo Study of Skewed Theta Distributions on DIF Indices.

The purpose of this study was to investigate the effects of skewed theta distributions on DIF indices. A comparison of Mantel Haenszel (Mantel & Haenszel, 1959, Lord's chi-square(Lord, 1980), and DFIT (Raju, van der Linden, & Fleer, in press) results are presented. Several questions are addressed:

(1) How does theta skewness effect the various DIF indices?

(2) Which method should be recommended for skewed theta distribution?

(3) Does the degree of skewness effect DIF indices?

(4) How does DFIT compare to the other methods with respect to skewness?

In an increasingly multicultural society, the study of bias is very important. Standardized tests are used to make high stakes decisions about college admissions to promotion decisions. Differential item functioning is a way of testing items to see if bias exists. There are many indices for studying DIF. Methods such as the traditional χ^2 approach (e.g. Scheueneman, 1979; Shepard et al.,1981), loglinear models (Mellenbergh, 1982), the Mantel-Haenszel χ^2

(Mantel & Haenszel, 1959; Holland & Thayer, 1988), standardization approaches (Dorans & Kulick, 1986; Dorans, 1989), logistic regression techniques (Swaminathan & Rogers, 1990), and the logistic discriminant function analysis (Miller, Spray, & Wilson, 1992) are all viable DIF detection methods. A relatively new method for assessing DIF is DFIT. DFIT is an important breakthrough in DIF analysis for two reasons. First, not only the individual items of a test but also the test as a whole can be assessed for DIF. Second, DFIT can be used with dichotomous and polytomous items. Polytomous item analysis is very important with the increased use of performance assessment in education.

The significance of the study is that in educational and psychological data the distributions one deals with are sometimes not normal. For example competency exams or licensure exams are skewed negatively and positively respectively. This in turn could translate into a skewed θ distribution. This study is very important for several reasons. First, the results will indicate if skew distributions influence DIF indices and to what degree. If the results indicate that there is no difference in DIF detection with these skew distributions then there is no need to investigate theta distributions in practice. However, if detection of DIF items is affected, practitioners must begin to look at the theta distributions carefully. The study will also shed light on which method

BEST COPY AVAILABLE

is the most robust with respect to skew. A recommendation will be made as to which method is best to use. This study examined the effects of skewed θ distributions on DIF indices to see if the DIF indices functioned as intended by their authors.

Method

Sample size

Monte Carlo procedures were used to generate the data under normal and skewed conditions. A two-parameter unidimensional dichotomous model was used. Data was simulated for 1,000 examinees and 500 examinees in each the focal and reference groups. These sample sizes should ensure adequate precision for parameter estimations prior to DIF/DTF analyses (Mislevy & Bock, 1990). A test length of 40 items was investigated. This test length is the most common in recent DIF studies (Fleer, 1993; Flowers, 1995; Raju, Oshima et al., 1995) and represents an average test length. 500 replications per cell was used for each simulated condition.

Ability distributions

The θ's for the examinees were drawn from a normal, positively, and negatively skewed distribution. The skewed distributions was simulated by a chi-square distribution with five degrees of freedom. Five degrees of freedom was

chosen to represent a high degree of skew. Lambert (1995) discussed the use of this skew as representative of management and productivity distributions. A chi-square with 15 degrees of freedom was used to simulate a moderate skew. Both negative and positive skews used the same degrees of freedom. These distributions were chosen to represent a large array of possible distributions and perhaps give insight into the effect of different levels of skew on the various DIF indices.

Item parameters

The item parameters came from Cohen and Kim, 1993. These parameters are representative of a wide range of test items. The *a* parameter ranges from 0.55 to 1.80 and the *b* parameter ranges from -0.04 to 2.46. These parameters have been used in many DIF studies(Fleer, 1993; Flowers, 1995; Raju, 1995).

Generation of DIF

Three proportions of test-wide unidirectional DIF(0%, 10%, and 20%) were simulated. For the 40 item test , 0, 4, and 8 items were imbedded with DIF. In addition, items were generated to simulate uniform DIF(for which $a_{iR}=a_{iF}$ and $b_{iR}\neq b_{iF}$) in all cases except the 20% DIF condition. The 20% condition had both uniform(items 5,10,15,20) and non-uniform(items 25,30,35,40) DIF items. Tables 1-3 represent the true item parameters for the 40 item test.

BEST COPY AVAILABLE

Programs

Item and ability parameters will be estimated by means of BILOG (Mislevy & Bock, 1990). The maximum marginal likelihood (MML) estimation procedure and EM algorithm were used to estimate the item parameters. Default values were used for all estimations. Lord's chi-square was computed using IRTDIF 1.0(Kim & Cohen, 1991). Mantel-Haenszel was computed using MHDIF (Raju,1988). DFIT was computed using DFITDU (Raju, 1995).

Parameter Equating Procedures

If parameter estimation is performed separately for two groups, the parameters of each group are not on the same scale. Therefore, one test must be linked with the other so values are on the same metric. The estimation of equating coefficients will be made by means of Baker's modified test characteristic curve method implemented by the EQUATE 2.1 computer program (Baker, 1993). All parameter estimates for the reference group were equated to the metrics of the focal group. Several researchers (Candell & Drasgow, 1988; Drasgow, 1987; Lautenschlager & Park, 1988; Lord, 1980; Miller & Oshima, 1992) have shown that an iterative linking procedure improves identification of biased items. To minimize error introduced by the equating procedure, a two-stage linking procedure was used in this study. Initial linking was done with all 40 items and a DIF analysis was

BEST COPY AVAILABLE

performed. Items that were identified as DIF were eliminated and the linking procedure was performed again. Finally, all items were transformed using the linking coefficients obtained by the second linking procedure. The two-stage linking was done separately for DFIT and Lord's chi-square.

Results/Conclusions

The effectiveness of each DIF index was determined by examining the true positive("hits") and the false positive rates.

Calculation of False Positive Proportions

Each table contains the actual number of items flagged as DIF and the corresponding proportion. The proportions for the false positives were calculated in the following manner. Each item is mutually exclusive with regard to category. An item cannot, for example, be a true positive and a false positive. Each index yielded an output matrix of 40(items) X 500(replications). Therefore, there are 20,000 total item values per cell. In the 0% DIF condition no DIF items were included. The proportion of false positives was calculated by dividing the number of flagged items by 20,000. In this condition any item that was flagged was a false positive. In the 10% DIF condition, 2,000 items were true DIF items(items 5,10,15,20 with 500 replications each). Therefore, the proportion of false positives in the 10% DIF conditions were calculated by the

BEST COPY AVAILABLE

number of flagged items(excluding items 5,10,15, and 20) divided by 18,000. 18,000 is the number of items excluding the DIF items(2,000). In the 20% DIF condition, 4,000 items were true DIF items(items 5,10,15,20,25,30,35,and 40 with 500 replications each). Therefore, the false positives in the 20% DIF conditions were calculated by the number of flagged items(excluding items 5,10,15,20,25,30,35,and 40) divided by 16,000. 16,000 is the number of items excluding the DIF items(4,000). The use of slightly different denominators insured that each item is counted in only one category(false or true positive).

Calculation of True Positive Proportions

The true positives are the "hits" in the matrix. For example, if an item is DIF a true positive means that item was flagged as such. The 0% DIF conditions had no DIF items therefore there were no true positives. The 10% DIF conditions had 2,000 DIF items(items 5,10,15,and 20 with 500 replications each). The true positive proportion was calculated by the number of those 2,000 items flagged divided by 2,000.

The 20% DIF conditions had 4,000 DIF items(items 5,10,15, 20,25,30,35,and 40 with 500 replications each). The true positive proportion was calculated by the number of those 4,000 items flagged divided by 4,000. In addition, this condition had both uniform and non-uniform DIF items. Each table for the 20% DIF condition contains a total true

BEST COPY AVAILABLE

positive proportion and also true positive proportions for uniform and non-uniform DIF items separately.

DFIT: Non-Compensatory DIF Results

Non-Compensatory DIF items were identified by using a cut-off value of .006 (a .01 level of significance). A two-stage linking procedure was performed and the items that were greater than .006 were eliminated and the DIF analysis was done again. NC-DIF detected uniform DIF items very well across all distributions. True positive rates ranged from .9990 to .9680 in the 10% DIF condition and .9998 to .9635 in the 20% DIF condition. Detection of Non-uniform DIF was substantially lower ranging from .4985 to .5730 in the 20% DIF condition. Sample size effected detection rates of both types of DIF. When sample size was increased to 1000 subjects, detection rates for uniform DIF approached unity. Non-uniform DIF detection was also improved but only slightly. False positive rates for all conditions approached the nominal alpha with ranges from .0006 to .0247. As sample size was increased, false positive rates decreased to negligible levels. Tables 4 and 5 contain the true and false positive rates for each condition.

Insert Tables 4 and 5 here

BEST COPY AVAILABLE

Lord's Chi-square Results

Lord's chi-square items were identified by using a chi-square with 1 degree of freedom at the .001 level of significance. McLaughlin and Drasgow (1987) suggest a nominal alpha of .001 in order to achieve an actual rejection rate of .01. The cut-off value of 10.827 was used. A two-stage linking procedure was performed and the items that were greater than the cut-off were eliminated and the DIF analysis was done again. Lord's chi-square detected uniform DIF items well across all distributions. True positive rates ranged from .9970 to .9475 in the 10% DIF condition and .9925 to .9280 in the 20% DIF condition. Detection of Non-uniform DIF was substantially lower ranging from .4130 to .4830 in the 20% DIF condition. Sample size effected detection rates of both types of DIF. When sample size was increased to 1000 subjects, detection rates for uniform DIF approached unity. Non-uniform DIF detection was improved. Detection rates went from the mid 40 percents to the low 70 percents. False positive rates for all conditions ranged from .0022 to .0086. Tables 6 and 7 contain the true and false positive rates for each condition.

Insert Tables 6 and 7 here

BEST COPY AVAILABLE

Mantel-Haenszel Results

Mantel-Haenszel items were identified by using a chi-square with 2 degree of freedom at the .01 level of significance. A cutoff value of 9.210 was utilized. Mantel-Haenszel detected uniform DIF items well across all distributions. True positive rates ranged from .9840 to .9025 in the 10% DIF condition and .9715 to .8305 in the 20% DIF condition. Detection of Non-uniform DIF was substantially lower ranging from .0615 to .3985 in the 20% DIF condition. Sample size effected detection rates of both types of DIF. When sample size was increased to 1000 subjects, detection rates for uniform DIF approached unity. Non-uniform DIF detection was slightly improved. As percent of DIF items increased, the number of false positives also increased. Tables 8 and 9 contain the true and false positive rates for each condition.

Insert Tables 8 and 9 here

DFIT: Compensatory DIF Results

Compensatory DIF items were identified by using a cut-off value of less than or equal to .006 or non significant chi-square was used with the DTF index for successive deletion of items. The proportions presented are in terms of identifying(deleting) true DIF items. C-DIF detected

BEST COPY AVAILABLE

uniform DIF items consistently across all distributions. However, this method was the least precise index of those studied. True positive rates ranged from .8055 to .7440 in the [illegible] DIF condition and .9460 to .8590 in the 20% DIF condition. Detection of Non-uniform DIF was substantially lower ranging from .2900 to .5840 in the 20% DIF condition. Sample size affected detection rates of both types of DIF. When sample size was increased to 1000 subjects, detection rates for uniform DIF ranged from .8590 to .9460. Non-uniform DIF detection was also improved but only slightly. False positive rates for all conditions approached the nominal alpha with ranges from .0026 to .0676. Tables 10 and 11 contain the true and false positive rates for each condition.

Insert Tables 10 and 11 here

Comparison of Indices

Mantel-Haenszel, Lord's chi-square, and NC-DIF detected uniform DIF items very well. Precision of detection increased as sample size increased for all methods. The three previously mentioned methods tended to have difficulty detecting non uniform DIF items. Previous research (Hambleton & Rogers, 1989; Hills, 1989; Clauser, Nungerster et al, 1994) suggests that Mantel-Haenszel is very weak at detecting non-uniform DIF and the empirical data of this study confirm this finding. Clauser, Nungester, et al (1994)

BEST COPY AVAILABLE

suggest that the extent to which the Mantel-Haenszel procedure is effective is related to the degree to which simulated DIF approximates uniform DIF across the curve representing most of the examinees in the sample, as was the case for our data.

NC-DIF weakness on identifying non uniform DIF items was an unexpected finding given that NC-DIF and C-DIF were designed to non uniform DIF items. One explanation maybe that the magnitude of actual DIF was not that large. Overall, C-DIF was the least stable measure. This discrepancy may be due to linking. This study utilized iterative linking on NC-DIF values. Possibly linking on C-DIF could raise the precision of detection. Also, DTF is calculated by summing C-DIF for each item. Any errors would thus be summed across items and could effect C-DIF detection rates.

Precision declines for all methods at the extreme skew. Moderate skew seems to have no effect. False positive rates were low for all methods. However, the number of false positives for Mantel-Haenszel, increased as the percent of DIF items increased. For example, false positive rates for the 0% DIF condition was .0018 but for the 20% DIF condition the false positive rate increased to .0705. This trend was exhibited in every condition.

How do the results of this study effect the practitioner? The results impact the practitioner in several ways. First, The presence of skewed distributions

BEST COPY AVAILABLE

does not seem to be as problematic as first thought. At the extreme skews precision of detection is slightly decreased, however, moderate skews do not effect detection much at all. All the methods functioned well with uniform DIF items and not very well with non-uniform DIF items. The data indicated that Mantel-Haenszel's false positive rates increased as the number of DIF items increased. More research in this area is needed to suggest why this trend occurred. Practitioners tend to be loyal to their DIF detection methods for various reasons. Some methods are simple to understand and easy to calculate, such as Mantel-Haenszel. Others use IRT principles, and with the capacity for computers increasing every day, are also easy to calculate. This study suggests that each of the methods did well at detection of DIF. Therefore, the recommendation is to use a combination of methods. For example, use Mantel-Haenszel and DFIT. Comparing results can led to more reliable decisions about items. The cost of making items is very high but the cost of implementing an addition DIF method is relatively low.

BEST COPY AVAILABLE

REFERENCES

Baker, F. B. (1993). EQUATE2.1: Computer program for equating two metrics in item response theory [computer program]. Madison: University of Wisconsin, Laboratory of Experimental Design.

Candell, G. L. & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. Applied Psychological Measurement, 12, 253-260.

Cohen, A. S. & Kim, S. H. (1993). A comparison of Lord's chi-square and Raju's area measures in detection of DIF. Applied Psychological Measurement, 17(1), 39-52.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. Applied Measurement in Education, 3, 217-233.

Dorans, N. J. & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude test. Journal of Educational Measurement, 23, 355-368.

Drasgow, F. (1987). Study of the measurement of bias of two standardized psychological tests. Journal of Applied Psychology, 72(1), 19-29.

BEST COPY AVAILABLE

Fleer, P. F. (1993). A monte carlo assessment of a new measure of item and test bias. Unpublished doctoral dissertation, Illinois Institute of Technology.

Flowers, C. P. (1995). A monte carlo assessment of DFIT with polytomously scored unidimensional tests. Unpublished doctoral dissertation, Georgia State University.

Holland, P.W., and Thayer, D.T.(1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer, and H. I. Braun (Eds.), Test validity (pp.129.145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lambert, R. G. (1995). The robustness of the standard error of summarized, corrected validity coefficients to non independent and non normality of primary data. Unpublished doctoral dissertation, Georgia State University.

Lautenshlager, G. J. & Park, D. G. (1988). IRT item bias detection procedures: Issues of model misspecification, robustness, and parameter linking. Applied Psychological Measurement, 12, 365-376.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Center, 22, 719-748.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. Journal of Educational Statistics, 7, 105-118.

BEST COPY AVAILABLE

Miller, M. D. & Oshima, T. C. (1992). Effect of sample size, number of biased items, and magnitude of bias on two-stage item bias estimation method. Applied Psychological Measurement, 16, 381-388.

Miller, T. R., Spray, J. A., & Wilson, A. (1992). A comparison of three methods for identifying non uniform DIF in polytomously scored test items. Paper presented at the Psychometric Society Meeting, Columbus, OH.

Mislevy, R. J., & Bock, R. D. (1990). BILOG3: Item and test scoring with binary logistic models. Mooresville, IN: Scientific Software.

Raju, N. S., van der Linden, W. J.,& and Fleer, P. F. (in press). An IRT-based internal measure of test bias with application for differential item functioning. Applied Psychological Measurement.

Raju, N. S. (1995). DFITDU: A fortran program for calculating DIF/DTF. [computer program]. Georgia Institute of Technology.

Raju, N. S. (1988). MHDIF: A fortran program for calculating Mantel-Haenszel statistic. [computer program]. Illinois Institute of Technology.

Scheuneman, J. D., & Bleistein, C. A. (1989). A consumer's guide to statistics for identifying differential item functioning. Applied Measurement in Education, 2, 255-275.

BEST COPY AVAILABLE

Shepard, L. A., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test item bias with both internal and external ability criteria. Journal of Statistics, 6, 317-375.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27, 361-370.

BEST COPY AVAILABLE

Table 1

Item Parameters for Generating 40-Item Test Conditions

Condition 1 (0% DIF)

	Reference		Focal	
Item	a	b	a	b
1	0.55	0.00		
2	0.55	0.00		
3	0.73	-1.04		
4	0.73	-1.04		
5	0.73	0.00		
6	0.73	0.00		
7	0.73	0.00		
8	0.73	0.00		
9	0.73	1.04		
10	0.73	1.04		
11	1.00	-1.96		
12	1.00	-1.96		
13	1.00	-1.04		
14	1.00	-1.04		
15	1.00	-1.04		
16	1.00	-1.04		
17	1.00	0.00		
18	1.00	0.00		
19	1.00	0.00		
20	1.00	0.00		
21	1.00	0.00		
22	1.00	0.00		
23	1.00	0.00		
24	1.00	0.00		
25	1.00	1.04		
26	1.00	1.04		
27	1.00	1.04		
28	1.00	1.04		
29	1.00	1.96		
30	1.00	1.96		
31	1.36	-1.04		
32	1.36	-1.04		
33	1.36	0.00		
34	1.36	0.00		
35	1.36	0.00		
36	1.36	0.00		
37	1.36	1.04		
38	1.36	1.04		
39	1.80	0.00		
40	1.80	0.00		

Note: Blanks indicate use of same item parameters listed for Reference group.

BEST COPY AVAILABLE

Table 2

Item Parameters for Generating 40-Item Test Conditions

Condition 2 (10% DIF)

	Reference		Focal	
Item	a	b	a	b
1	0.55	0.00		
2	0.55	0.00		
3	0.73	-1.04		
4	0.73	-1.04		
5	0.73	0.00	0.73	1.00
6	0.73	0.00		
7	0.73	0.00		
8	0.73	0.00		
9	0.73	1.04		
10	0.73	1.04	0.73	2.04
11	1.00	-1.96		
12	1.00	-1.96		
13	1.00	-1.04		
14	1.00	-1.04		
15	1.00	-1.04	1.00	-0.04
16	1.00	-1.04		
17	1.00	0.00		
18	1.00	0.00		
19	1.00	0.00		
20	1.00	0.00	1.00	0.50
21	1.00	0.00		
22	1.00	0.00		
23	1.00	0.00		
24	1.00	0.00		
25	1.00	1.04		
26	1.00	1.04		
27	1.00	1.04		
28	1.00	1.04		
29	1.00	1.96		
30	1.00	1.96		
31	1.36	-1.04		
32	1.36	-1.04		
33	1.36	0.00		
34	1.36	0.00		
35	1.36	0.00		
36	1.36	0.00		
37	1.36	1.04		
38	1.36	1.04		
39	1.80	0.00		
40	1.80	0.00		

Note: Blanks indicate use of same item parameters listed for Reference group.

Condition 2 contains 4 uniform DIF items (5,10,15,20) favoring the Reference group.

BEST COPY AVAILABLE

Table 3

Item Parameters for Generating 40-Item Test Conditions

Condition 3 (20% DIF)

	Reference		Focal	
Item	a	b	a	b
1	0.55	0.00		
2	0.55	0.00		
3	0.73	-1.04		
4	0.73	-1.04		
5	0.73	0.00	0.73	1.00
6	0.73	0.00		
7	0.73	0.00		
8	0.73	0.00		
9	0.73	1.04		
10	0.73	1.04	0.73	2.04
11	1.00	-1.96		
12	1.00	-1.96		
13	1.00	-1.04		
14	1.00	-1.04		
15	1.00	-1.04	0.50	-0.54
16	1.00	-1.04		
17	1.00	0.00		
18	1.00	0.00		
19	1.00	0.00		
20	1.00	0.00	0.50	0.0
21	1.00	0.00		
22	1.00	0.00		
23	1.00	0.00		
24	1.00	0.00		
25	1.00	1.04	1.00	2.04
26	1.00	1.04		
27	1.00	1.04		
28	1.00	1.04		
29	1.00	1.96		
30	1.00	1.96	1.00	2.46
31	1.36	-1.04		
32	1.36	-1.04		
33	1.36	0.00		
34	1.36	0.00		
35	1.36	0.00	0.86	0.50
36	1.36	0.00		
37	1.36	1.04		
38	1.36	1.04		
39	1.80	0.00		
40	1.80	0.00	1.30	0.00

Note: Blanks indicate use of same item parameters listed for Reference group.

Condition 3 contains 4 uniform DIF items (5,10,25,30) and 4 non-uniform items (15,20,35,40).

BEST COPY AVAILABLE

Table 4 DFIT: NC-DIF Results

True Positives(Hits)

condition	normal	positive high	positive low	negative high	negative low
500,0	-----	-----	-----	-----	-----
500,10	.9985	.9705	.9990	.9680	.9985
500,20	.7820	.7325	.7857	.7572	.7522
uniform	.9980	.9665	.9998	.9635	.9965
non-uniform	.5660	.4985	.5730	.5510	.5080
1000,0	-----	-----	-----	-----	-----
1000,10	1.00	.9930	1.00	.9650	1.00
1000,20	.7765	.7492	.7732	.7705	.7572
uniform	.9995	.9885	1.00	.9810	1.00
non-uniform	.5535	.5100	.5465	.5600	.5145

Table 5

False Positives for DFIT: NC-DIF

condition	normal	positive high	positive low	negative high	negative low
500,0	.0006	.0101	.0146	.0104	.0148
500,10	.0168	.0130	.0163	.0130	.0161
500,20	.0216	.0159	.0216	.0165	.0247
1000,0	.0002	.0002	.0004	.0003	.0002
1000,10	.0004	.0003	.0005	.0028	.0005
1000,20	.0009	.0004	.0007	.0006	.0006

BEST COPY AVAILABLE

Table 6 Lord's chi-square results

True Positives(Hits)

condition	normal	positive high	positive low	negative high	negative low
500,0	-----	-----	-----	-----	-----
500,10	.9920	.9515	.9875	.9475	.9970
500,20	.7342	.6920	.7370	.6865	.7012
uniform	.9925	.9480	.9920	.9280	.9895
non-uniform	.4760	.4350	.4830	.4450	.4130
1000,0	-----	-----	-----	-----	-----
1000,10	1.00	.9995	1.00	.9965	1.00
1000,20	.8750	.8392	.8752	.8927	.8442
uniform	1.00	.9990	1.00	.9993	1.00
non-uniform	.7500	.6795	.7505	.7925	.6885

Table 7

False Positives for Lord's chi-square

condition	normal	positive high	positive low	negative high	negative low
500,0	.0032	.0028	.0024	.0024	.0025
500,10	.0028	.0030	.0025	.0030	.0022
500,20	.0031	.0040	.0040	.0040	.0048
1000,0	.0027	.0032	.0030	.0031	.0029
1000,10	.0033	.0036	.0045	.0041	.0053
1000,20	.0063	.0041	.0045	.0080	.0086

BEST COPY AVAILABLE

Table 8 Mantel-Haenszel Results

True Positives(Hits)

condition	normal	positive high	positive low	negative high	negative low
500,0	-----	-----	-----	-----	-----
500,10	.9775	.9025	.9710	.8930	.9840
500,20	.5977	.4747	.5805	.6145	.6077
uniform	.9715	.8880	.9665	.8305	.9680
non-uniform	.2240	.0615	.1940	.3985	.2475
1000,0	-----	-----	-----	-----	-----
1000,10	1.00	.9995	1.00	.9985	1.00
1000,20	.6982	.5952	.6830	.6635	.6945
uniform	.9995	.9905	.9995	.9530	1.00
non-uniform	.3970	.2000	.3670	.3740	.3890

Table 9

False Positives for Mantel-Haenszel

condition	normal	positive high	positive low	negative high	negative low
500,0	.0020	.0116	.0016	.0013	.0013
500,10	.0118	.0084	.0013	.0012	.0143
500,20	.0218	.0118	.0195	.0610	.0253
1000,0	.0018	.0016	.0016	.0014	.0015
1000,10	.0370	.0202	.0371	.0360	.0389
1000,20	.0705	.0301	.0606	.1619	.0806

BEST COPY AVAILABLE

Table 10 DFIT: C-DIF Results

True Positives(Hits)

condition	normal	positive high	positive low	negative high	negative low
500,0	-----	-----	-----	-----	-----
500,10	.8055	.7440	.7940	.7450	.7955
500,20	.5507	.4990	.5360	.6597	.5050
uniform	.7305	.7045	.7365	.7355	.7075
non-uniform	.3710	.2900	.3370	.5840	.3025
1000,0	-----	-----	-----	-----	-----
1000,10	.9460	.9140	.9455	.8590	.9350
1000,20	.6575	.4111	.6611	.7470	.6297
uniform	.8245	.8115	.8230	.8365	.8150
non-uniform	.4905	.0105	.4990	.6570	.4445

Table 11

False Positives for DFIT: C-DIF

condition	normal	positive high	positive low	negative high	negative low
500,0	.0108	.0212	.0116	.0216	.0115
500,10	.0230	.0233	.0211	.0316	.0217
500,20	.0211	.0299	.0158	.0236	.0198
1000,0	.0026	.0069	.0050	.0162	.0052
1000,10	.0122	.0137	.0142	.0176	.0148
1000,20	.0136	.0676	.0110	.0137	.0136

BEST COPY AVAILABLE